TU Dortmund

Design of Computer Experiments and Active Learning

Active learning: Minimising expected error and variance

Lecturers:
JProf. Dr. Kirsten Schorning

Author: Sathish Kabatkar Ravindranth

Matrik Nr: 229208

July 31, 2023

Contents

1 Introduction **1**

2 Problem description **1**

3 Statistical methods **2**

3.1 Assumptions 2

3.2 Minimising the expected error 2

3.3 Minimising the expected log-loss 3

3.4 Minimising the variance in quadratic loss 4

3.5 Variance reduction using optimal experimental design 5

3.6 Variance reduction with submodular function in batch learning 6

4 Application **7**

4.1 Datasets 7

4.2 Algorithms implemented 8

5 Conclusion **10**

Bibliography **11**

1 Introduction

Active learning is a specific type of machine learning where a learning algorithm has the capability to interactively ask a user or another source of information to provide labels for new data points. Unlike traditional learning, which relies on pre-labeled data, active learning actively formulates hypotheses and seeks new information through targeted queries. By selecting informative instances for labeling, active learners efficiently make inferences with less training data, reducing the burden of annotating vast amounts of unlabeled data. The goal of minimizing expected errors and variance in predictions has led to various query selection frameworks and utility measures. Active learning encompasses three fundamental querying strategies: uncertainty sampling, query-by-committee, and stream-based selective sampling, each enriching the machine's decision-making capabilities from different angles (Settles, 2022, p. 4).

The aim of this seminar report is to explore the fundamental principles of active learning querying by minimising expected error and variance, and analysing its profound significance with an example and discussing the reduced labeling costs and augmenting the efficacy of machine learning models across real-world applications.

In Section 2, a detailed description of the problem statement is discussed. In Section 3, statistical methods are explained, such as minimising the expected error with log-loss and output variance in quadratic loss using the bias-variance trade-off and fisher score. In Section 4, the application of this querying strategy is explained and the results of the method are discussed with plots. Finally, the central results and discussions of the seminar report are summarized in Section 5.

2 Problem description

The main focus lies in binary classification, with an initial setup of a classification model. The objective is to enhance the predictive accuracy of this model using active learning. The process involves selecting specific instances to query and adding them to the training data to improve the model's performance. However, determining which instances are the most suitable to query requires careful consideration. Although some approaches rely on model certainty or hypothesis correctness, there are scenarios where predictive accuracy becomes more critical. The challenge lies in the uncertainty surrounding the answers to the queries before asking them and the model's error even after updating

its hypotheses. Consequently, decision-making under uncertainty becomes a significant aspect of the problem (Settles, 2022, p. 37).

3 Statistical methods

3.1 Assumptions

Consider an oracle, which provides a true label y, for a query instance x. To estimate the expected error, two probability distributions are necessary. The first is the probability distribution of the true label y, which the oracle provides in response to a query instance x. Once the oracle's label known, the second distribution needed is the probability of the learner or the fitted classification model making an error on a different instance $x^{'}$. The model's posterior distribution can be utilized as an approximation for these assumptions.

3.2 Minimising the expected error

In decision theory, the expected error is used for decision-making under uncertainty. Rather than seeking to know exact errors, this approach focuses on calculating expected values based on the probabilities of various outcomes. By considering all possible outcomes and their probabilities to make more informed decisions and minimize the impact of uncertainties. Consider the unlabeled data $\mathcal{U}$, the optimisation problem to choose an instance x^* which minimises the expected classification error is given by the formula:

$$\begin{aligned} x^*_{ER} &= \underset{x}{\operatorname{argmin}}\ \mathbb{E}_{Y|\theta,x}\left[\sum_{x'\in\mathcal{U}} \mathbb{E}_{Y|\theta^+,x'}[y \neq \hat{y}]\right] \\ &= \underset{x}{\operatorname{argmin}} \sum_{y} P_\theta(y \mid x)\left[\sum_{x'\in\mathcal{U}} 1 - p_{\theta^+}\left(\hat{y} \mid x'\right)\right]. \end{aligned}$$

The formula represents a key aspect of active learning, aiming to identify the most informative instance, denoted as x^*_{ER}, for querying in order to minimize the expected error. Once its true label y is known given the current model parameters θ and the instance itself x to select the instance that is most likely to reduce the overall error in the future.

The term $\mathbb{E}_{Y|\theta,x}\left[\sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta^+,x'}[y \neq \hat{y}]\right]$ represents the expected error. The term $y \neq \hat{y}$ represents the error, where y is the true label and $\hat{y}$ is the model's predicted label for a given instance. Here, given the current model parameters θ and the instance x, the outer expectation is taken over the true label Y. The inner summation considers all instances $x^{'}$ in the unlabeled pool $\mathcal{U}$, and given the updated model parameters θ^+ and each instance $x^{'}$, the inner expectation is taken over the true label Y .

The expected error is further simplified. The term $\sum_y P_\theta(y \mid x)$ represents the probability of the true label Y given the instance x and the current model parameters θ. The subsequent summation over instances $x^{'}$ in the unlabeled pool $\mathcal{U}$ involves the term $p_{\theta^+}(\hat{y} \mid x')$ denotes the probability of the model's predicted label $\hat{y}$ for each instance $x^{'}$ after updating the model parameters θ^+ (Settles, 2022, p. 38).

3.3 Minimising the expected log-loss

Minimizing the expected error directly is challenging due to uncertainty in the true labels and model predictions. Instead, minimizing the log loss using entropy as it quantifies the uncertainty in a probability distribution and the optimisation problem to choose an instance x^* which minimises the expected log loss is given by the formula:

$$
\begin{aligned}
x^*_{LL} &= \underset{x}{\operatorname{argmin}}\, \mathbb{E}_{Y|\theta,x}\left[\sum_{x' \in \mathcal{U}} \mathbb{E}_{Y|\theta^+,x'}\left[-\log p_{\theta^+}\left(y \mid x'\right)\right]\right] \\
&= \underset{x}{\operatorname{argmin}} \sum_y P_\theta(y \mid x)\left[\sum_{x' \in \mathcal{U}} -\sum_{y'} p_{\theta^+}\left(y' \mid x'\right) \log p_{\theta^+}\left(y' \mid x'\right)\right] \\
&= \underset{x}{\operatorname{argmin}} \sum_y P_\theta(y \mid x) \sum_{x' \in \mathcal{U}} H_{\theta^+}\left(Y \mid x'\right),
\end{aligned}
$$

To identify the instance x^*_{LL} that minimizes the expected log loss, where log loss measures the discrepancy between true labels and model predictions. The expected log loss is calculated by taking the expectations over true labels Y given the current model parameters θ and the instance x. Further the log loss is simplified by considering the probability of model predictions for each instance $x^{'}$ in the unlabeled pool $\mathcal{U}$ after updating the model parameters θ^+. The sum of the log losses for all outcomes measures the uncertainty or entropy associated with the model's predictions for that instance $x^{'}$. The entropy, represented as H, is a measure of the randomness or unpredictability in a given set of probabilities. The entropy $H_{\theta^+}(Y \mid x^{'})$ calculates the uncertainty in the

model's predicted labels for each instance $x^{'}$ in the unlabeled pool $\mathcal{U}$ after updating the model parameters θ^{+} (Settles, 2022, p. 38).

3.4 Minimising the variance in quadratic loss

Minimizing error functions directly can be problematic due to complexity and processing costs, while estimating predicted error reduction can be time-consuming. However, indirectly reducing output variance in the quadratic loss is an effective way to improve model performance and achieve more accurate predictions. This simplifies optimization and proves valuable in regression problems aiming to minimize standard error. Geman et al. (1992) demonstrated the decomposition of expected error in the quadratic loss into:

$$\begin{aligned}\mathbb{E}\left[(\hat{y}-y)^2 \mid x\right] &= \underbrace{\mathbb{E}_{Y|x}\left[\left(y-\mathbb{E}_{Y|x}[y \mid x]\right)^2\right]}_{\text{noise}} + \underbrace{\left(\mathbb{E}_{\mathcal{L}}[\hat{y}]-\mathbb{E}_{Y|x}[y \mid x]\right)^2}_{\text{bias}} \\ &\quad + \underbrace{\mathbb{E}_{\mathcal{L}}\left[(\hat{y}-\mathbb{E}_{\mathcal{L}}[\hat{y}])^2\right]}_{\text{output variance}}.\end{aligned}$$

The formula is a valuable insight into the decomposition of expected error for a regression problem aiming to minimize standard error (squared-loss). The expected error, denoted by $\mathbb{E}\left[(\hat{y}-y)^2 \mid x\right]$, is expressed as a sum of three distinct components.

The "noise" component captures the uncertainty or variability in the true label y given the instance x. It reflects the inherent randomness in the relationship between the instance and its true label, which contributes to the overall error in predictions. The "bias" component measures the deviation between the expected model prediction $\mathbb{E}_{\mathcal{L}}[\hat{y}]$ and the true label's expectation $\mathbb{E}_{Y|x}[y \mid x]$. It indicates how much the model's predictions deviate, on average, from the true values. The "output variance" component quantifies the variability or spread of the model's predictions $\hat{y}$ around their average $\mathbb{E}_{\mathcal{L}}[\hat{y}]$ over the labeled training set $\mathcal{L}$.

Minimizing variance is a valuable strategy to reduce future generalization error in the model, especially when the learner has limited control over noise or bias components. To reduce the model output variance in the unlabeled data $\mathcal{U}$, the strategy involves choosing instances that can possibly reduce:

$$x^{*}_{VR} = \underset{x}{\operatorname{argmin}} \sum_{x' \in \mathcal{U}} \operatorname{Var}_{\theta^{+}}\left(Y \mid x'\right)$$

where θ^+ denotes the model after re-training.

3.5 Variance reduction using optimal experimental design

The primary focus lies in regression tasks involving a limited number of input variables as predictors. A prominent approach to obtain closed-form utility functions for active learning is through "Optimal Experimental Design," as introduced by (Chaloner and Verdinelli, 1995) and (Fedorov, 1972). At the core of this method is the concept of Fisher information, which "quantifies the information carried by a random variable Y about a parameter θ in the likelihood function $P_\theta(Y)$". The fisher score, denoted as ∇x, is a crucial element calculated as "the partial derivative of the logarithm of $P_\theta(Y \mid x)$ with respect to θ", where x represents an input instance, and Y depends on it. By utilizing the Fisher score, the amount of information gained by labeling a specific instance during the active learning process is given by the formula:

$$\nabla x = \frac{\partial}{\partial \theta} \log P_\theta(Y \mid x).$$

The fisher score, denoted as ∇x (or gradient), relies solely on the distribution over Y under parameters θ and is independent of the actual label of instance x. In cases involving multiple parameters, the fisher score is represented as a vector and can be calculated as the expected value of the squared partial derivative of the logarithm of $P_\theta(Y \mid x)$ with respect to θ.

$$F = \mathbb{E}_X \left[\left(\frac{\partial}{\partial \theta} \log P_\theta(Y \mid x) \right)^2 \right] \quad = \mathbb{E}_X \left[\frac{\partial^2}{\partial \theta^2} \log P_\theta(Y \mid x) \right] \propto \sum_x \nabla x \nabla x^\top \quad .$$

The fisher information, denoted as F, corresponds to the variance of the fisher score. It is not dependent on a specific observation since it involves integrating over all instances. Alternatively, to explicitly indicate their relationship with model parameters, the fisher score is represented as $\nabla_\theta x$ and information is represented as $F(\theta)$.

In active learning, choosing instance that minimizes the inverse fisher information or maximizes the fisher information is essential to reduce variance in parameter estimates. For models with K parameters, Fisher information is represented as a K x K covariance matrix, which complicates the optimization process. There are two main optimal experimental designs: D-optimality and A-optimality. D-optimality aims to minimize

the differential posterior entropy of parameter estimates and is expressed as:

$$x_D^* = \arg\min_x \mathbf{det}\left(\left[F_{\mathcal{L}} + \nabla x \nabla x^\top\right]^{-1}\right)$$

where ∇x represents the fisher score of the input instance x, and $F_{\mathcal{L}}$ is the fisher information matrix based on the labeled set $\mathcal{L}$ (Chaloner and Verdinelli, 1995). Conversely, A-optimal designs prioritize elements along the main diagonal of fisher information matrix for minimizing the average variance of parameter estimates. A simple variant is to minimize $\mathbf{tr}\left(AF_{\mathcal{L}}^{-1}\right)$, where A is a square, symmetric "reference" matrix. A-optimal designs attempt to minimize the prediction variance for all data instances. Using A-optimal design, the utility measure for the optimisation problem is derived as:

$$\begin{aligned} x_{FIR}^* = \operatorname*{argmin}_x \sum_{x' \in \mathcal{U}} \mathrm{Var}_{\theta^+}\left(Y \mid x'\right) &= \operatorname*{argmin}_x \sum_{x' \in \mathcal{U}} \mathbf{tr}\left(A_{x'}\left[F_{\mathcal{L}} + \nabla x \nabla x^\top\right]^{-1}\right) \\ &= \operatorname*{argmin}_x \mathbf{tr}\left(F_{\mathcal{U}}\left[F_{\mathcal{L}} + \nabla x \nabla x^\top\right]^{-1}\right). \end{aligned}$$

where $\mathrm{Var}_{\theta^+}(.)$ denotes the variance of the re-trained model with query x and its label y.

3.6 Variance reduction with submodular function in batch learning

Active learning often involves selecting queries one at a time, but batch-mode active learning is a variation that selects queries in groups. This makes it suitable for scenarios with parallel labeling or slow training processes. Variance reduction heuristics are well-suited for batch-mode active learning and offer performance guarantees. Submodularity is a property that captures the concept of diminishing returns when adding elements to a set (Nemhauser et al., 1978). In active learning, variance reduction heuristics can be formulated as submodular functions, ensuring performance guarantees. The utility function used for variance reduction quantifies the change in the model's output variance before and after selecting a set of queries $\mathcal{Q}$. Greedy algorithms based on submodular criteria can efficiently select queries in certain learning algorithms like Gaussian processes, logistic regression, and linear regression.

The submodularity property of a set function s is captured by the inequalities:

$$s(\mathcal{A} \cup \{x\}) - s(\mathcal{A}) \geq s\left(\mathcal{A}' \cup \{x\}\right) - s\left(\mathcal{A}'\right)$$

or

$$s(\mathcal{A}) + s(\mathcal{B}) \geq s(\mathcal{A} \cup \mathcal{B}) + s(\mathcal{A} \cap \mathcal{B})$$

where $\mathcal{A} \subseteq \mathcal{A}'$ and $\mathcal{A}, \mathcal{A}', \mathcal{B}$ are sets. The utility function for variance reduction can be expressed as:

$$s(\mathcal{Q}) = \sum_{x \in \mathcal{U}} \mathrm{Var}_{\theta}(Y \mid x) - \mathrm{Var}_{\theta^{+}\mathcal{Q}}(Y \mid x) = \mathrm{tr}\left(F_{\mathcal{U}} F_{\mathcal{L}}^{-1}\right) - \mathrm{tr}\left(F_{\mathcal{U}}\left[F_{\mathcal{L}} + F_{\mathcal{Q}}\right]^{-1}\right)$$

where $\mathcal{U}$ is the unlabeled pool, $\mathcal{L}$ is the labeled pool, $\mathcal{Q}$ is the set of queried instances, θ and θ^{+} represent model parameters before and after querying, respectively, and F denotes the fisher information matrix.

4 Application

The application of error-reduction sampling in the active learning field has demonstrated promising results, particularly in scenarios where obtaining labeled data is expensive or time-consuming. This approach was applied to a dataset to evaluate its performance relative to other active learning algorithms.

4.1 Datasets

The study utilized a dataset for active learning experiments (Roy and McCallum, 2001). The data used in the study was Ken Lang's Newsgroups, 20 different UseNet discussion groups across 20,000 articles. The first experiment objective was binary classification, specifically distinguishing between "comp.graphics" and "comp.windows.x" classes. Before analyzing the data, several pre-processing procedures were conducted, which involved eliminating UseNet headers and UUencoded binary data, converting alphabetic sequences into words, and excluding stoplist words and infrequent terms. These steps led to a final vocabulary of 10,205 words. The second experiment involved a more challenging text-categorization problem, where the goal was to classify newsgroups "comp.sys.ibm.pc.hardware" and "comp.os.ms-windows.misc." Preprocessing was similar to the first experiment, resulting in a vocabulary of 9,895 words. Both experiments

began with six labeled examples, and at each iteration, 250 documents were randomly sampled from the unlabeled data pool for candidate labeling (Ken, 2008).

4.2 Algorithms implemented

The initial classification model trained is a Naive bayes classifier. The four different active learning algorithms were tested: Random, Uncertainty Sampling, Density-Weighted QBC, and Error-Reduction Sampling. Initially six labeled instances, three from each class is trained with the algorithms. For labeling the candidates, 250 documents (25% unlabeled documents) are randomly sampled.

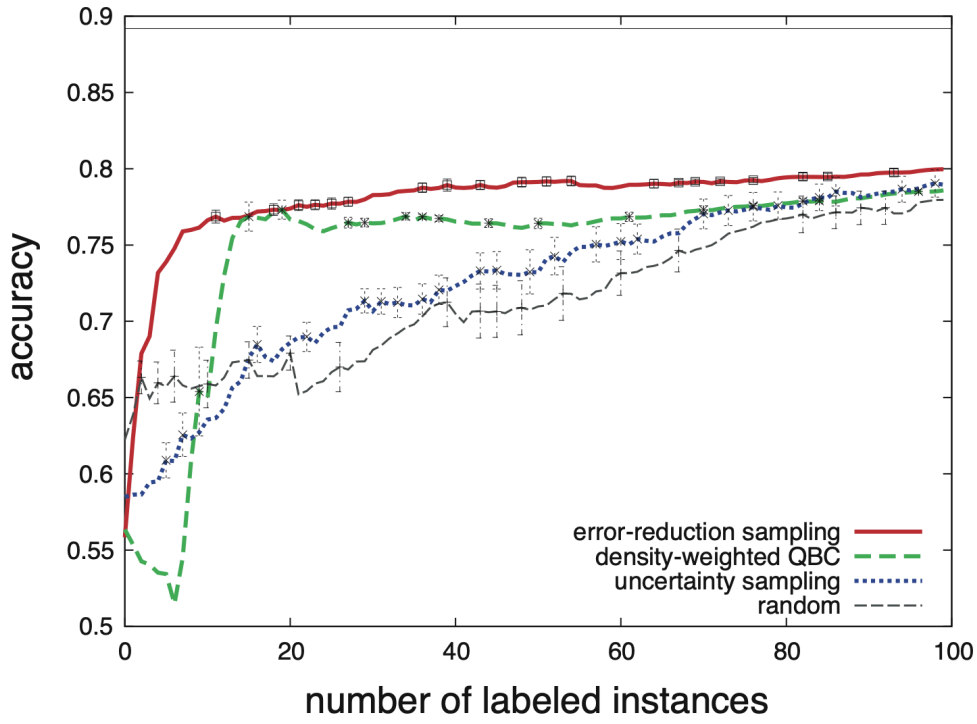

Figure 1: Average test set accuracy for comp.graphics vs. comp.windows.x..

Figure 1 illustrates the active learning process to differentiate between the newsgroups "comp.graphics" and "comp.windows.x.". The classification accuracy on test set is given on the vertical axis and and the horizontal axis denotes the number of queries performed (up to 100). The results are averaged over 10 trials. When all unlabeled data is labeled, the maximum accuracy achieved is 89.2% marked as a solid line. The Error-Reduction Sampling algorithm showed rapid progress, achieving 77.2% accuracy in just 16 queries. In contrast, the Density-Weighted QBC algorithm took 68 queries to reach the same

accuracy level, making it four times slower, with lower accuracy for the remaining queries. An intriguing observation relates to the documents selected for initial labeling by the two algorithms. The Error-Reduction Sampling algorithm consistently favored informative documents like FAQs, tutorials, or HOW-TO guides in 9.8 out of 10 cases during the first 10 queries. In contrast, the Density-Weighted QBC algorithm chose such documents only 5.8 times out of 10. Though not precisely quantified, this suggests that Error-Reduction Sampling exhibits more intuitive behavior in the initial stages of active learning.

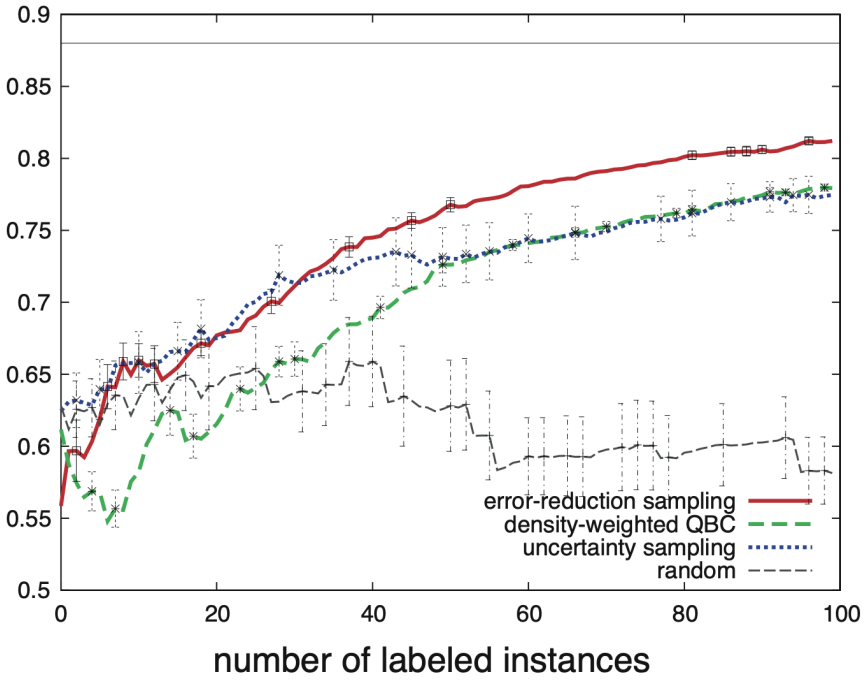

Figure 2: Average test set accuracy for the comp.sys.ibm.pc.hardware vs. comp.os.ms-windows.misc.

Figure 2 illustrates the results of a challenging text-categorization experiment that aimed to differentiate between the newsgroups "comp.sys.ibm.pc.hardware" and "comp.os.ms-windows.misc." These categories were deliberately selected for their complexity compared to the previous experiment. After applying similar preprocessing techniques, the data resulted in a vocabulary of 9,895 words. Once again, the dataset was divided into two separate sets, with 1000 documents allocated for training and the remaining 1000 documents designated for testing. Like before, 250 unlabeled documents were randomly sampled at each iteration for potential labeling, and the sampling error was evaluated against all remaining unlabeled documents. In the initial stages, Error-Reduction Sam-

pling predominantly chose 7.3 frequently asked questions (FAQs) out of the first 10 selected documents, while Density-Weighted QBC only had an average of 2.6 FAQs. Surprisingly, this intuitive behavior did not lead to significant superiority of one algorithm over the other. Both approaches required additional labeled documents to attain reasonable accuracy.

When all unlabeled data is labeled, the maximum accuracy achieved is 88% marked as a solid line. To achieve 75% accuracy, the error-reduction technique took 42 queries which was 1.6 times faster than the Density weighted QBC method which took 70 queries.

5 Conclusion

This seminar report explored systematic approaches to active learning that focus on minimizing expected error or output variance for unlabeled data points. These methods have demonstrated potential in creating more precise models while using fewer labeled instances compared to uncertainty-based approaches. Although strategies focused on classification error can be computationally intensive, variance reduction heuristics provide more efficient closed-form calculations. Moreover, these techniques can be applied to both classification and regression problems.

Active learners using expected error and variance reduction heuristics select instances that effectively reduce model output variance, leading to faster convergence to high accuracy with limited labeled data. However, these approaches may not be as widely applicable as uncertainty-based methods and can be computationally expensive for large-scale problems with high-dimensional feature spaces.

Despite their limitations, expected error and variance reduction methods provide valuable insights and powerful tools for active learning. They offer a principled approach to optimize data labeling and have demonstrated their potential in various real-world applications.

In conclusion, expected error or variance reduction-based active learning strategies hold promise for advancing machine learning by minimizing expected error and output variance and improving model performance with limited labeled instances. Further research and innovation in variance reduction approaches can pave the way for more efficient and effective machine learning systems in the future.

Bibliography

Chaloner, K. and Verdinelli, I. (1995), *Bayesian experimental design: A review. Statistical Science, 10(3):237–304. DOI: 10.1214/ss/1177009939.*

Fedorov, V. V. (1972), *Theory of optimal experiments. Neural Computation, 4:1–58, DOI: 10.1162/neco.1992.4.1.1*, Academic press.

Geman, S., Bienenstock, E. and Doursat, R. (1992), *Neural networks and the bias/variance dilemma. Neural Computation, 4:1–58, DOI: 10.1162/neco.1992.4.1.1.*

Ken, L. (2008), *20 Newsgroup original, URL: https://www.kaggle.com/datasets/au1206/20-newsgroup-original (visited on 31st July 2023).*

Nemhauser, G., Wolsey, L. and Fisher, M. (1978), *An analysis of approximations for maximizing submodular set functions. Mathematical Programming, 14(1):265–294. DOI: 10.1007/BF01588971.*

Roy, N. and McCallum, A. (2001), *Toward Optimal Active Learning through Sampling Estimation of Error Reduction , In Proceedings of the International Conference on Machine Learning (ICML), pages 441–448. Morgan Kaufmann (visited on 31st July 2023).*

Settles, B. (2022), *Active learning*, Springer. ISBN 978-3-031-01560-1.